#3

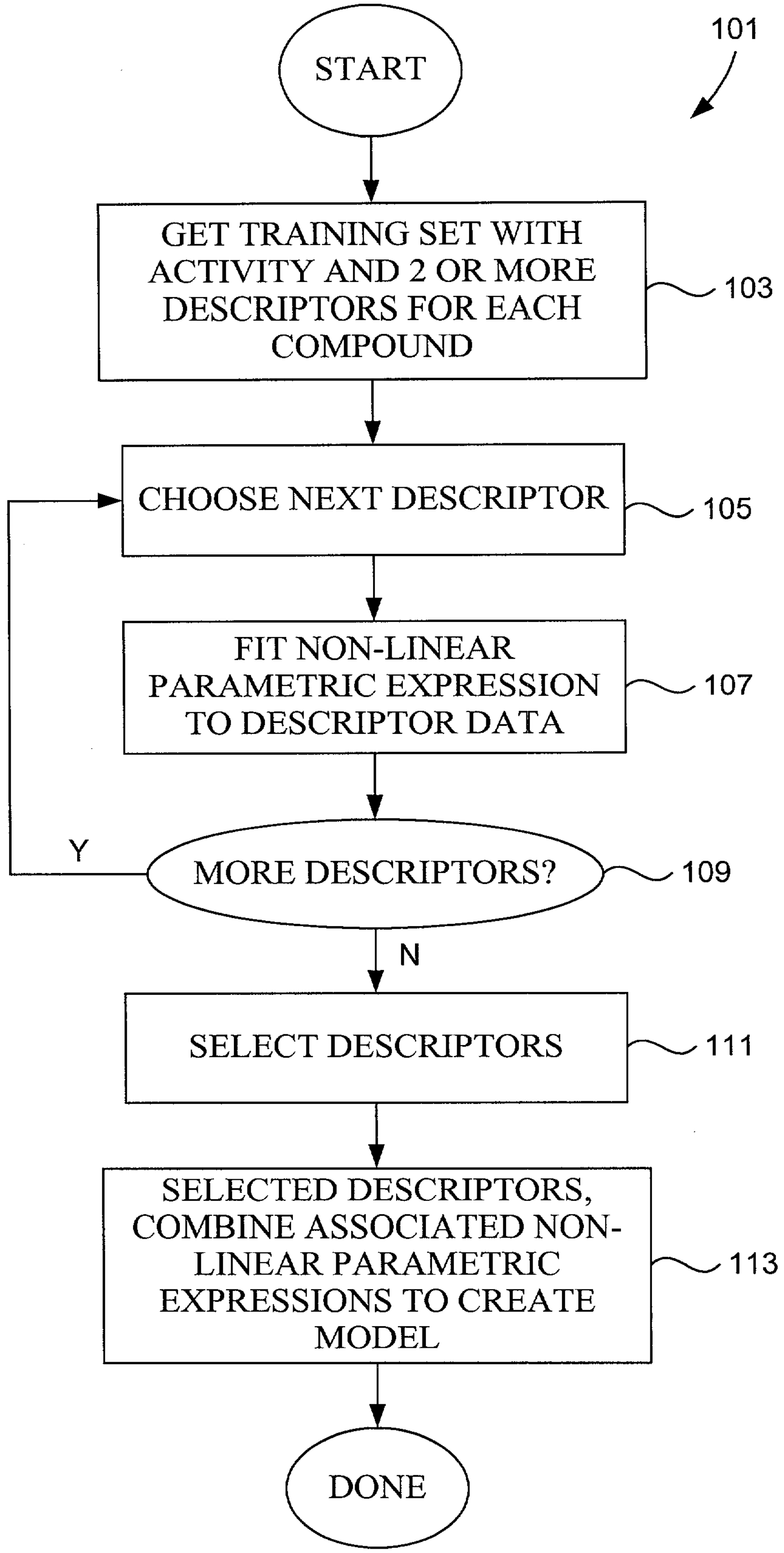

101
START
GET TRAINING SET WITH ACTIVITY AND 2 OR MORE DESCRIPTORS FOR EACH COMPOUND
103
CHOOSE NEXT DESCRIPTOR
105
FIT NON-LINEAR PARAMETRIC EXPRESSION TO DESCRIPTOR DATA
107
Y
MORE DESCRIPTORS?
109
N
SELECT DESCRIPTORS
111
SELECTED DESCRIPTORS, COMBINE ASSOCIATED NON-LINEAR PARAMETRIC EXPRESSIONS TO CREATE MODEL
113
DONE

FIGURE 1

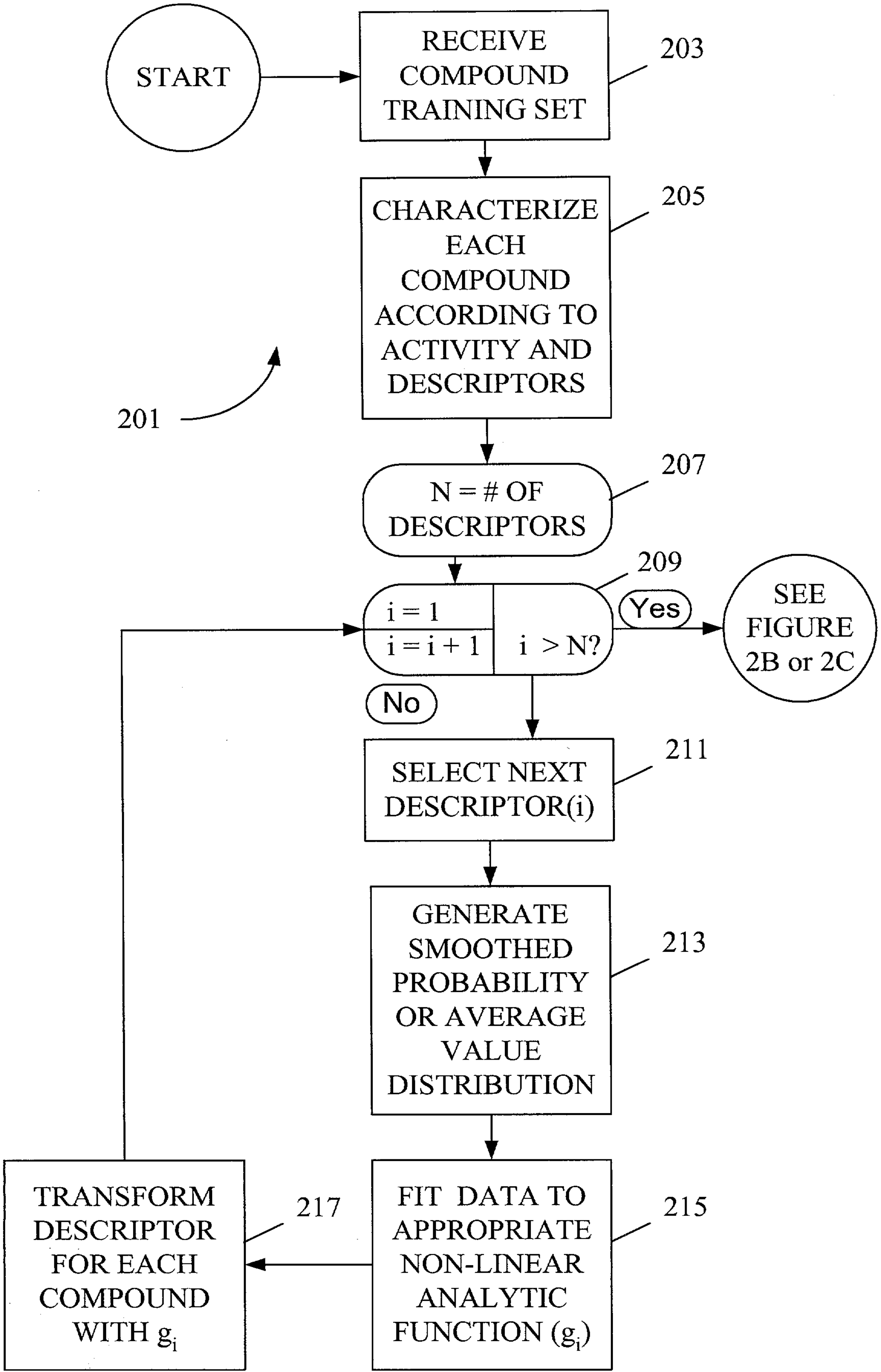

START
RECEIVE COMPOUND TRAINING SET
203
CHARACTERIZE EACH COMPOUND ACCORDING TO ACTIVITY AND DESCRIPTORS
205
201
N = # OF DESCRIPTORS
207
209
i = 1
i = i + 1
i > N?
Yes
SEE FIGURE 2B or 2C
No
SELECT NEXT DESCRIPTOR(i)
211
GENERATE SMOOTHED PROBABILITY OR AVERAGE VALUE DISTRIBUTION
213
FIT DATA TO APPROPRIATE NON-LINEAR ANALYTIC FUNCTION (g_i)
215
TRANSFORM DESCRIPTOR FOR EACH COMPOUND WITH g_i
217

FIGURE 2A

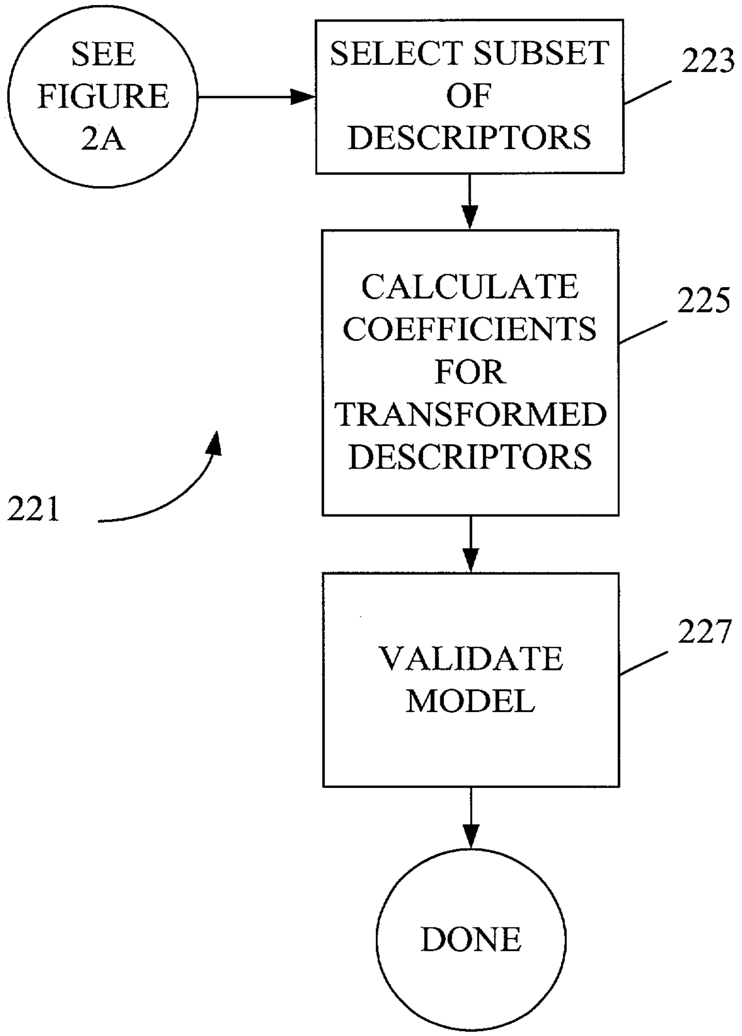

SEE FIGURE 2A
SELECT SUBSET OF DESCRIPTORS
223
CALCULATE COEFFICIENTS FOR TRANSFORMED DESCRIPTORS
225
221
VALIDATE MODEL
227
DONE

FIGURE 2B

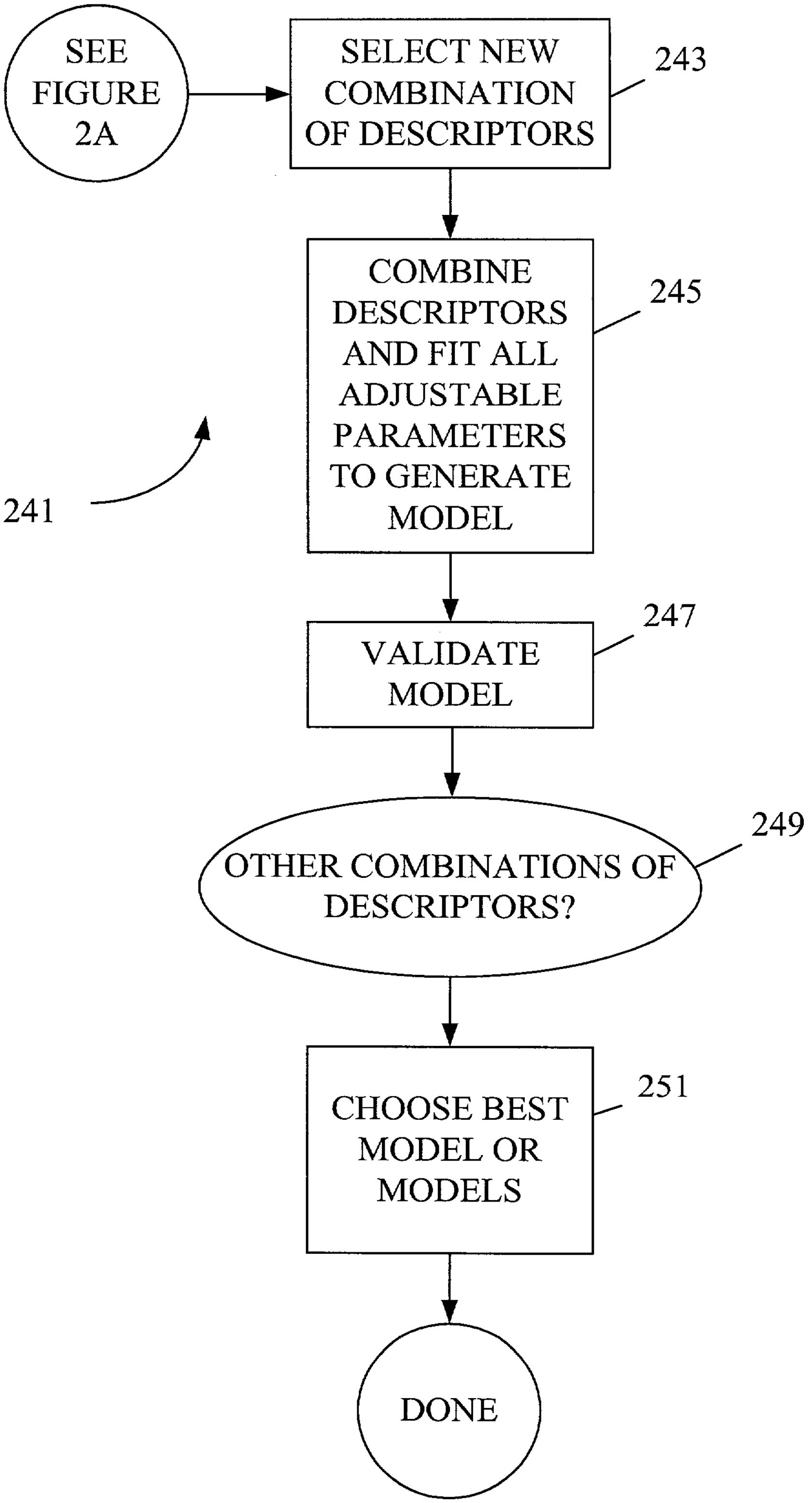

SEE FIGURE 2A
SELECT NEW COMBINATION OF DESCRIPTORS
243
COMBINE DESCRIPTORS AND FIT ALL ADJUSTABLE PARAMETERS TO GENERATE MODEL
245
241
VALIDATE MODEL
247
OTHER COMBINATIONS OF DESCRIPTORS?
249
CHOOSE BEST MODEL OR MODELS
251
DONE

FIGURE 2C

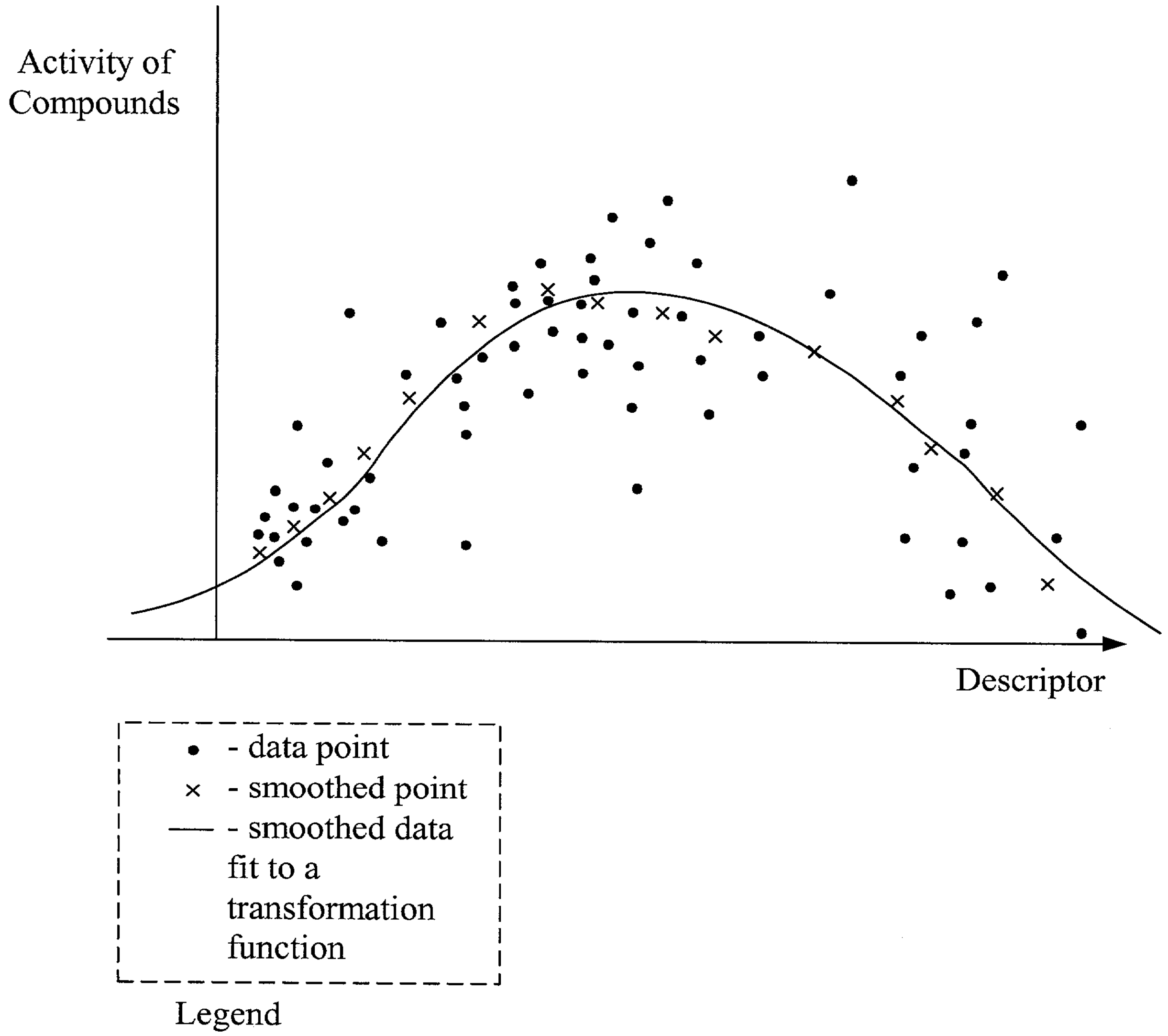

Activity of Compounds
Descriptor
• - data point
× - smoothed point
— - smoothed data fit to a transformation function
Legend

FIGURE 3

Optimum Molecular Weight

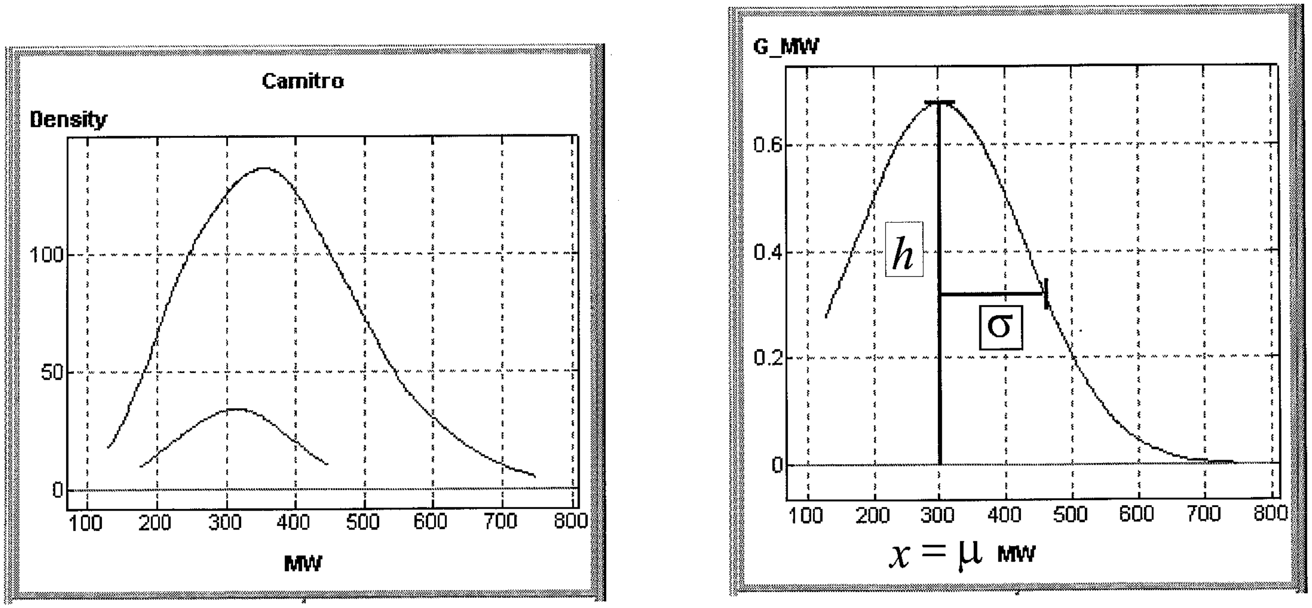


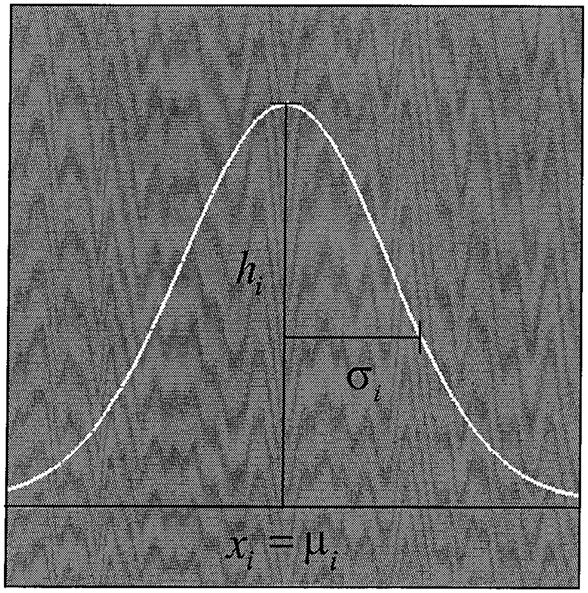
FIGURE 4A

$$g(x) = h e^{-(x-\mu)^2 / 4\sigma^2}$$

N-Dimensional Gaussian Modeling

Additive

$$g(x_{i..N}) = \frac{1}{N}\sum_{i=1}^{N} h_i\, e^{-(x_i-\mu_i)^2 / 4\sigma_i^2}$$



Multiplicative

$$g(x_{i..N}) = h\, e^{-\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_i)^2 / 4\sigma_i^2}$$

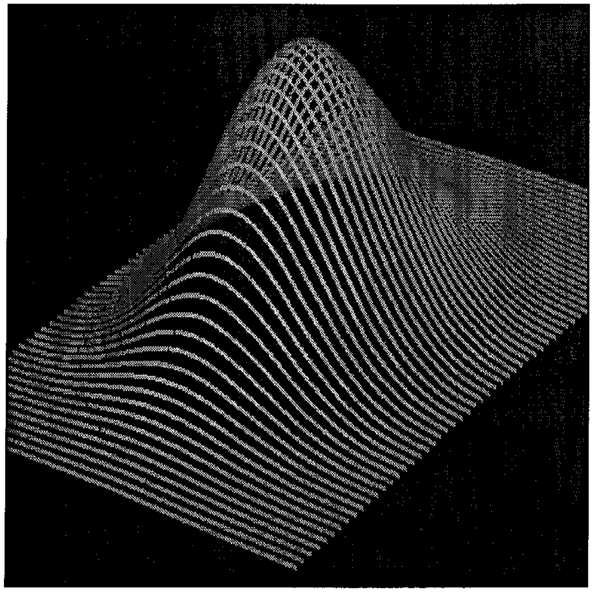


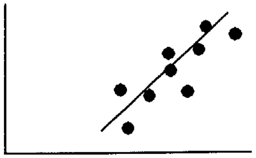
FIGURE 4B

Optimization Function

$$\mathbf{g}\,(\mathbf{x}, \mu, \sigma, h, t) = t + h\,e^{-\sum_{k=1}^{N_x} (x_k - \mu_k)^2 / 4\sigma_k^{\,2}}$$

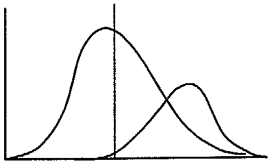
$$\boldsymbol{f} = s_{inh}\left[\frac{1}{N_{inh}}\sum_{i=1}^{N_{inh}}\left(\mathbf{g}\,(\mathbf{X}_i, \mu, \sigma, h, t) - y_i\right)^2\right]$$

Mean of the Squared Errors of Inhibitor Affinity



$$+\ s_{drug}\left[\frac{1}{N_{drug}}\sum_{j=1}^{N_{drug}}\mathbf{g}\,(\mathbf{X}_j, \mu, \sigma, h, t) - \bar{y}_{drug}\right]^2$$

Squared Error of the Means of Drug Affinity



$$+\ s_{fit}\left[\sigma_y^2\sum_{k=1}^{N_x}\left(\frac{\mu_k - \mu_{0,k}}{\boldsymbol{range}(\mathbf{X}_k^T)}\right)^2 + (t - t_0)^2\right]$$

Constraints to prevent Overfitting

FIGURE 4C

Initial Values for Optimization

$$t_0 = \boldsymbol{min}(\mathbf{y})$$

$$h_0 = \boldsymbol{max}(\mathbf{y}) - t_0$$

$$\mu_{0,k} = \frac{\sum_{i=1}^{N_{inh}} \left(y_i - t_0\right)^2 x_{k,i}}{\sum_{i=1}^{N_{inh}} \left(y_i - t_0\right)^2}$$

$$\sigma_{0,k} = \sqrt{\frac{\sum_{i=1}^{N_{inh}} \left(y_i - t_0\right)^2 \left(x_{k,i} - \mu_k\right)^2}{\sum_{i=1}^{N_{inh}} \left(y_i - t_0\right)^2}}$$

FIGURE 4D

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{N_{inh}} \left(y_i - \bar{y}_{inh}\right)^2}{N_{inh} - 1}}$$

Gaussian Optimization Function

$$f(\mathbf{x}) = t + h e^{-\sum_{k=1}^{N_x} (x_k - c_k)^2 / 4{w_k}^2}$$

$$f_o \;=\; s_y \left(\frac{\sum_{i=1}^{N_y} u_i \left(f(\mathbf{x}_i) - y_i \right)^2}{\sigma_y^2 \sum_{i=1}^{N_y} u_i} \right)$$

Weighted Mean Squared Error

$$+ \quad s_c \sum_{k=1}^{N_x} \frac{\left(c_k - c_{0,k} \right)^2}{\sigma_{x_k}^2}$$

Center Constraint

$$+ \quad s_w \sum_{k=1}^{N_x} \frac{\sigma_{x_k}}{w_k}$$

Width (Focus) Constraint

$$+ \quad s_t \frac{\left(t - t_0 \right)^2}{\sigma_y^2}$$

Tare Constraint

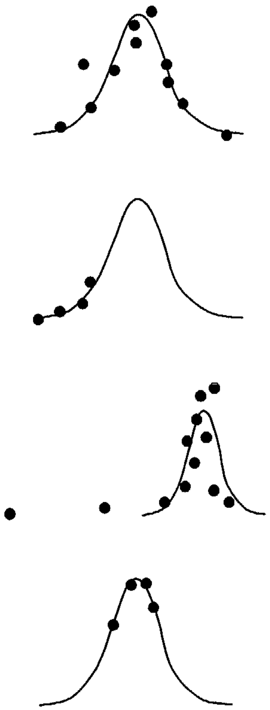


FIGURE 4E

Gaussian Optimization Starting Values

$$\sigma_y^2 = \frac{\sum_{i=1}^{N_{inh}} u_i \left(y_i - \bar{y}\right)^2}{\sum_{i=1}^{N_{inh}} u_i}$$

$$\sigma_{x_k}^2 = \frac{\sum_{i=1}^{N_{inh}} u_i \left(x_{k,i} - \bar{x}_k\right)^2}{\sum_{i=1}^{N_{inh}} u_i}$$

$$t_0 = \boldsymbol{min}(\mathbf{y})$$

$$h_0 = \boldsymbol{max}(\mathbf{y}) - t_0$$

$$v_i = \frac{\left(y_i - t_0\right)^2}{\sigma_y^2}$$

$$c_{0,k} = \frac{\sum_{i=1}^{N_{inh}} u_i v_i x_{k,i}}{\sum_{i=1}^{N_{inh}} u_i v_i}$$

$$w_{0,k}^2 = \frac{\sum_{i=1}^{N_{inh}} u_i v_i \left(x_{k,i} - c_{0,k}\right)^2}{\sum_{i=1}^{N_{inh}} u_i v_i}$$

FIGURE 4F

Performance Metrics

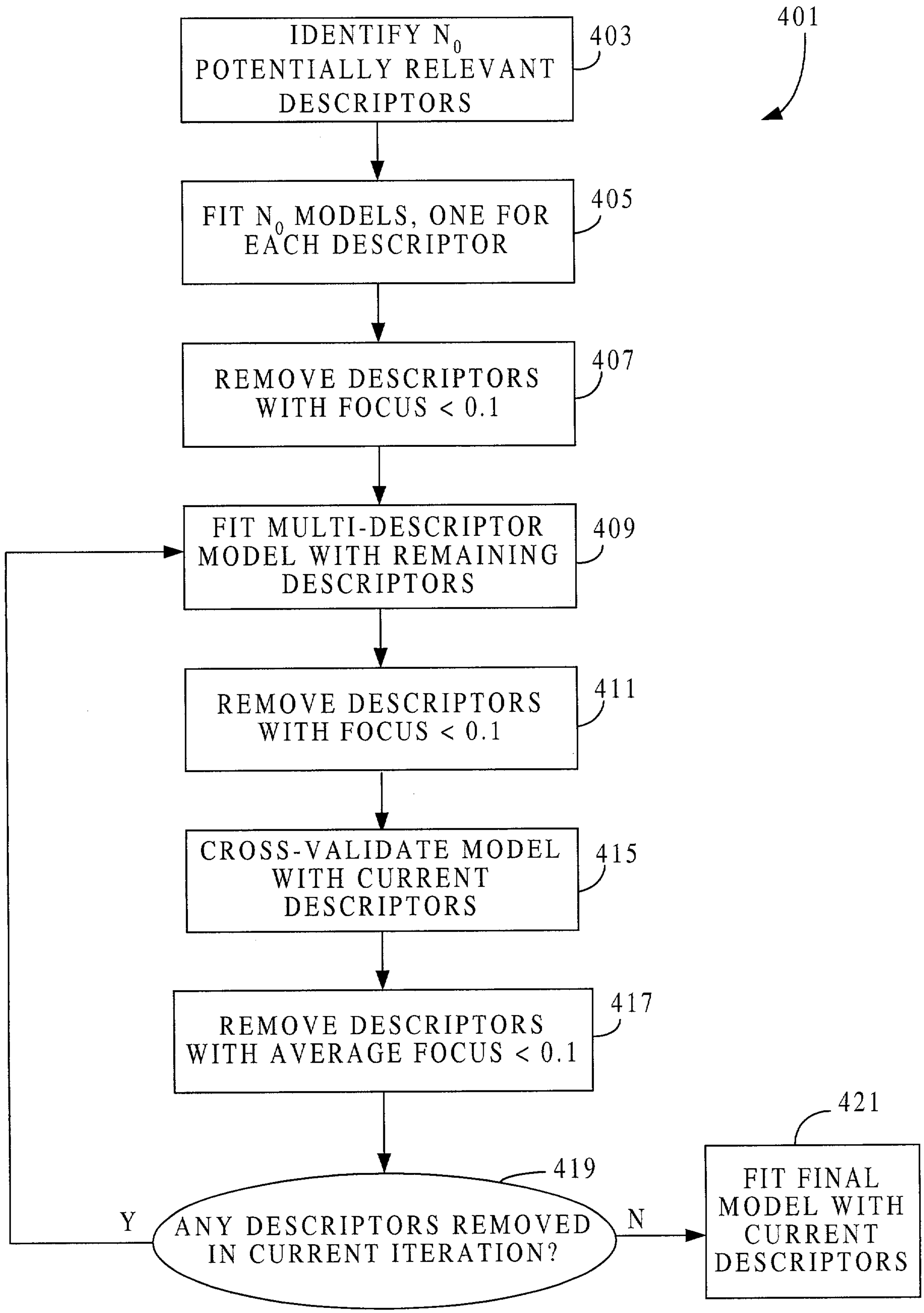
$$n_k = \frac{\sigma_{x_k}}{w_k}$$ Descriptor Focus

$$s = \sqrt{\frac{\sum_{i=1}^{N_y} u_i \left(f(\mathbf{x}_i) - y_i \right)^2}{\sum_{i=1}^{N_y} u_i}}$$ Standard Error

$$r^2 = \frac{\left(\sum_{i=1}^{N_y} u_i \left(f(\mathbf{x}_i) - \bar{f}(\mathbf{x}) \right) \left(y_i - \bar{y} \right) \right)^2}{\sum_{i=1}^{N_y} u_i \left(f(\mathbf{x}_i) - \bar{f}(\mathbf{x}) \right)^2 \sum_{i=1}^{N_y} u_i \left(y_i - \bar{y} \right)^2}$$ Correlation Coefficient

$$q^2 = 1 - s^2 / \sigma_y^2$$ Residual Error

FIGURE 4G



401
IDENTIFY N_0 POTENTIALLY RELEVANT DESCRIPTORS
403
FIT N_0 MODELS, ONE FOR EACH DESCRIPTOR
405
REMOVE DESCRIPTORS WITH FOCUS < 0.1
407
FIT MULTI-DESCRIPTOR MODEL WITH REMAINING DESCRIPTORS
409
REMOVE DESCRIPTORS WITH FOCUS < 0.1
411
CROSS-VALIDATE MODEL WITH CURRENT DESCRIPTORS
415
REMOVE DESCRIPTORS WITH AVERAGE FOCUS < 0.1
417
ANY DESCRIPTORS REMOVED IN CURRENT ITERATION?
419
Y
N
FIT FINAL MODEL WITH CURRENT DESCRIPTORS
421

FIGURE 4H

Sigmoid Optimization Function

$$f(\mathbf{x}) = t + \frac{h}{1 + \sum_{k=1}^{N_x} e^{-n_k (x_k - c_k)}}$$

$$\begin{aligned}
f_o \quad = \quad & s_y \left(\frac{\sum_{i=1}^{N_y} u_i \left(f(\mathbf{x}_i) - y_i \right)^2}{\sigma_y^2 \sum_{i=1}^{N_y} u_i} \right) && \text{Weighted Mean Squared Error} \\
+ \quad & s_c \sum_{k=1}^{N_x} \frac{\left(c_k - c_{0,k} \right)^2}{\sigma_{x_k}^2} && \text{Center Constraint} \\
+ \quad & s_n \sum_{k=1}^{N_x} \left| n_k \right| \sigma_{x_k} && \text{Focus Constraint} \\
+ \quad & s_t \frac{\left(t - t_0 \right)^2}{\sigma_y^2} && \text{Tare Constraint}
\end{aligned}$$

FIGURE 4I

Sigmoid Optimization Starting Values

$$t_0 = \boldsymbol{min(}\,\mathbf{y}\,\boldsymbol{)}$$

$$h_0 = \boldsymbol{max(}\,\mathbf{y}\,\boldsymbol{)} - t_0$$

$$v_i = \frac{(y_i - t_0)^2}{\sigma_y^2}$$

$$v_i' = \frac{(h_0 + t_0 - y_i)^2}{\sigma_y^2}$$

$$c_{h,k} = \frac{\sum_{i=1}^{N_{inh}} u_i v_i x_{k,i}}{\sum_{i=1}^{N_{inh}} u_i v_i}$$

$$c_{l,k} = \frac{\sum_{i=1}^{N_{inh}} u_i v_i' x_{k,i}}{\sum_{i=1}^{N_{inh}} u_i v_i'}$$

$$c_{0,k} = \frac{c_{h,k} + c_{l,k}}{2}$$

$$w_{h,k} = \sqrt{\frac{\sum_{i=1}^{N_{inh}} u_i v_i \left(x_{k,i} - c_{h,k}\right)^2}{\sum_{i=1}^{N_{inh}} u_i v_i}}$$

$$w_{l,k} = \sqrt{\frac{\sum_{i=1}^{N_{inh}} u_i v_i' \left(x_{k,i} - c_{l,k}\right)^2}{\sum_{i=1}^{N_{inh}} u_i v_i'}}$$

$$n_{0,k} = \frac{c_{h,k} - c_{l,k}}{w_{h,k} w_{l,k}}$$

FIGURE 4J

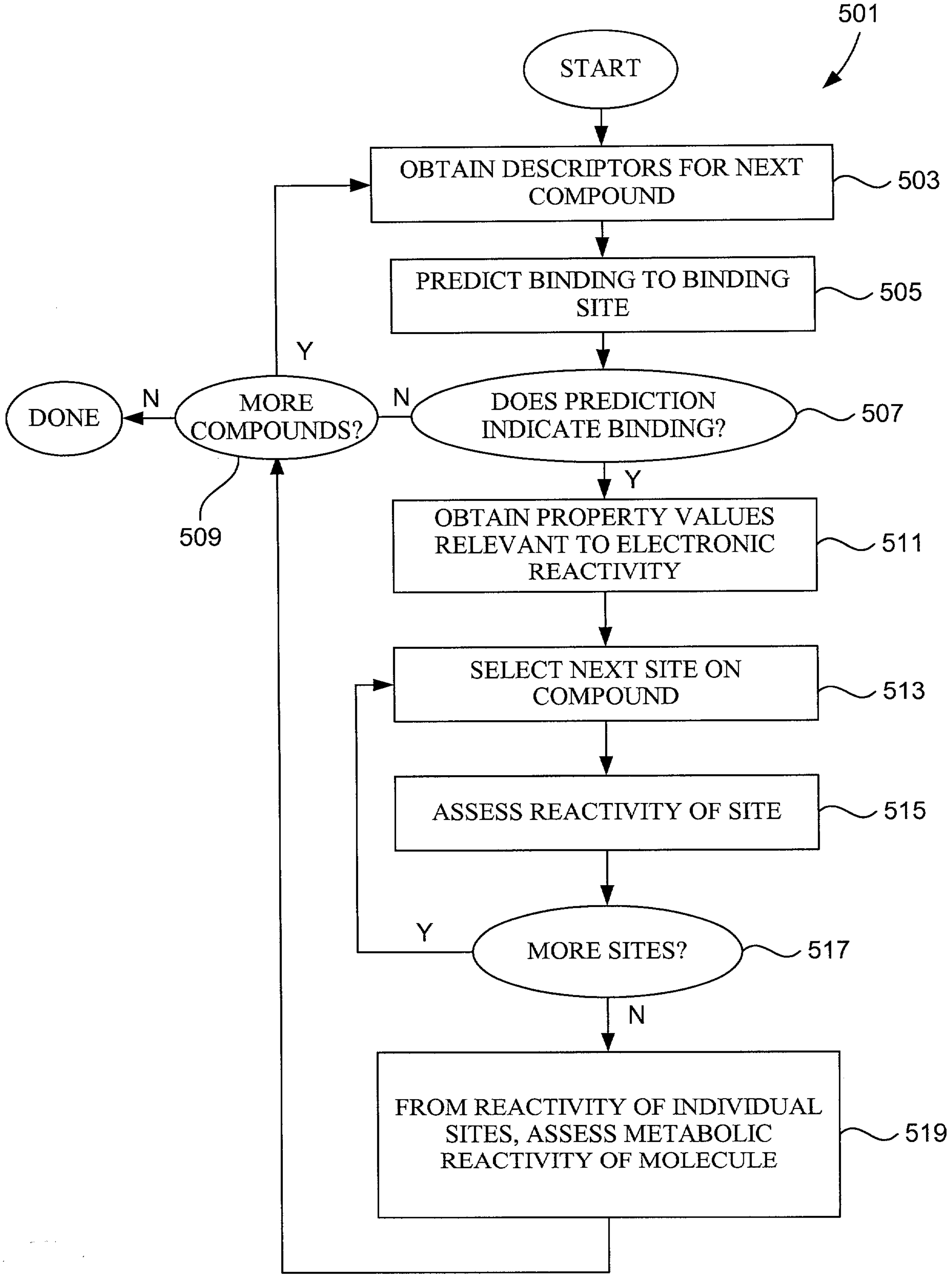

501
START
OBTAIN DESCRIPTORS FOR NEXT COMPOUND
503
PREDICT BINDING TO BINDING SITE
505
DOES PREDICTION INDICATE BINDING?
507
N
Y
MORE COMPOUNDS?
509
Y
N
DONE
OBTAIN PROPERTY VALUES RELEVANT TO ELECTRONIC REACTIVITY
511
SELECT NEXT SITE ON COMPOUND
513
ASSESS REACTIVITY OF SITE
515
MORE SITES?
517
Y
N
FROM REACTIVITY OF INDIVIDUAL SITES, ASSESS METABOLIC REACTIVITY OF MOLECULE
519

FIGURE 5

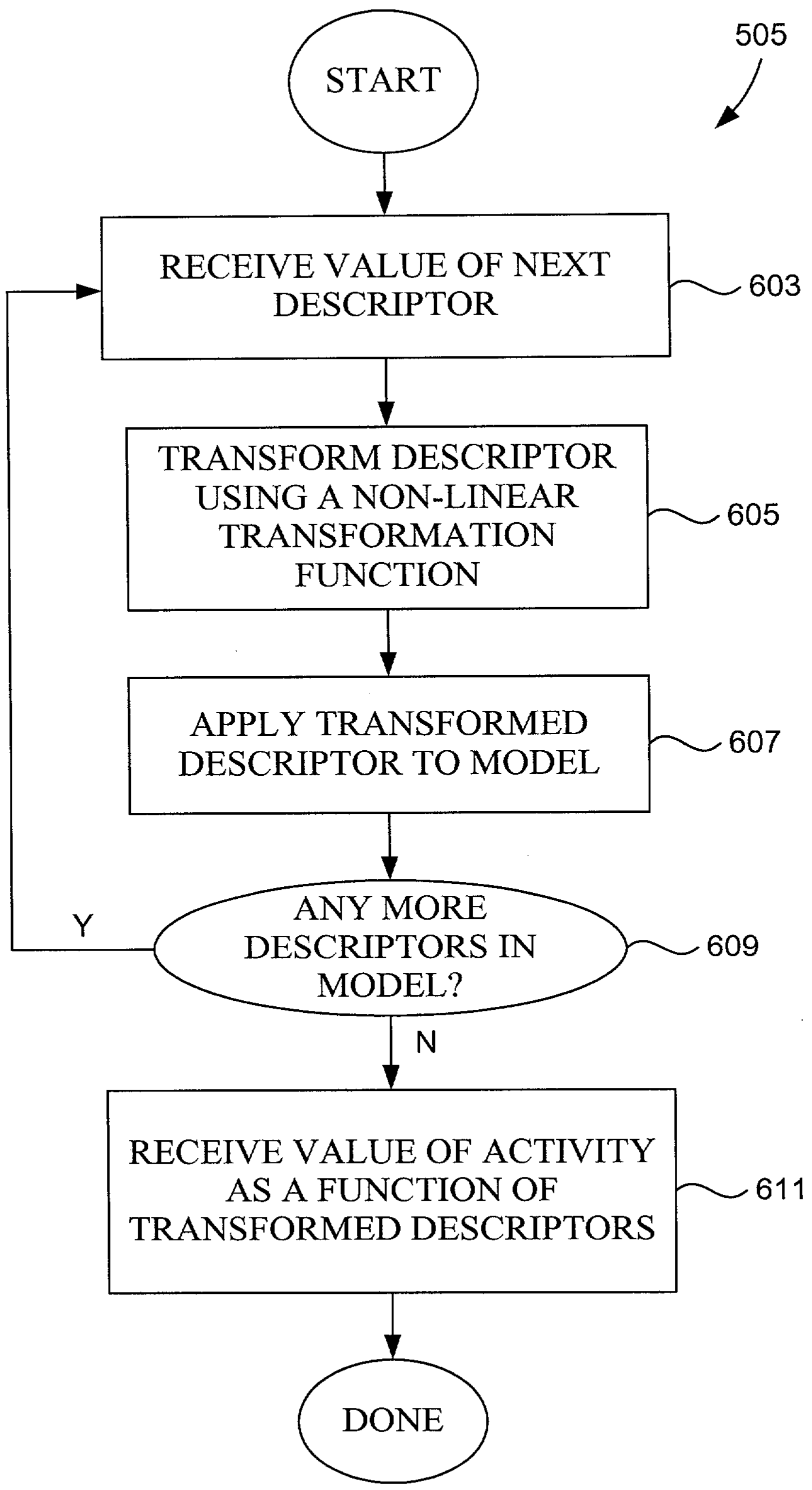

505
START
RECEIVE VALUE OF NEXT DESCRIPTOR
603
TRANSFORM DESCRIPTOR USING A NON-LINEAR TRANSFORMATION FUNCTION
605
APPLY TRANSFORMED DESCRIPTOR TO MODEL
607
ANY MORE DESCRIPTORS IN MODEL?
Y
609
N
RECEIVE VALUE OF ACTIVITY AS A FUNCTION OF TRANSFORMED DESCRIPTORS
611
DONE

FIGURE 6

Optimum logP

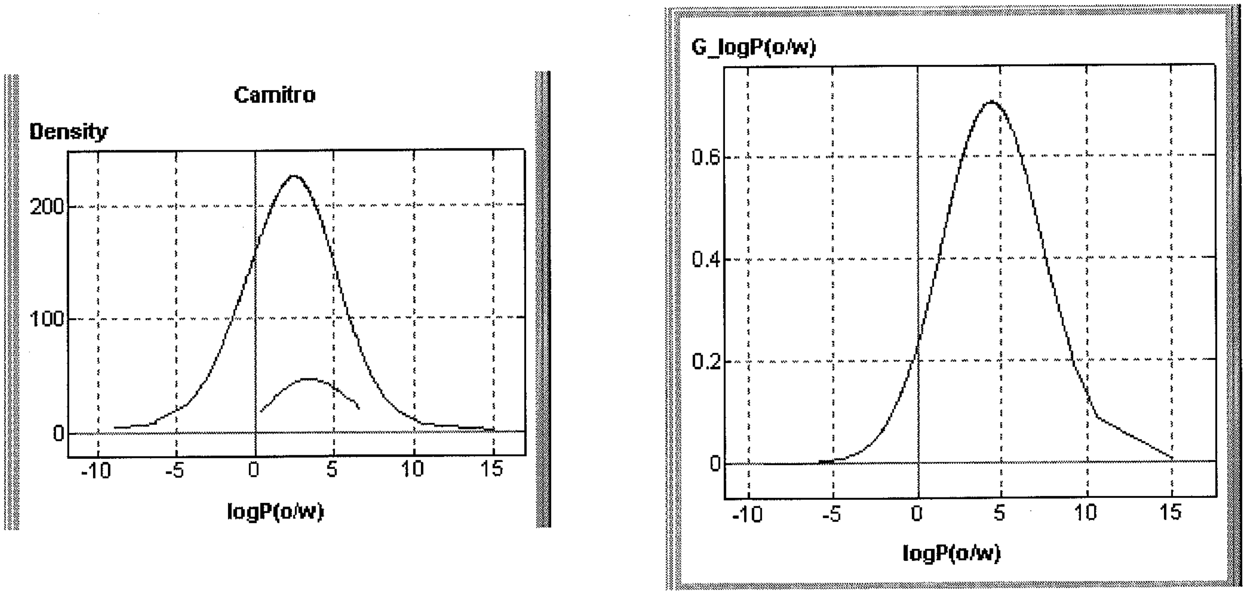


FIGURE 7A

Optimum Formal Charge

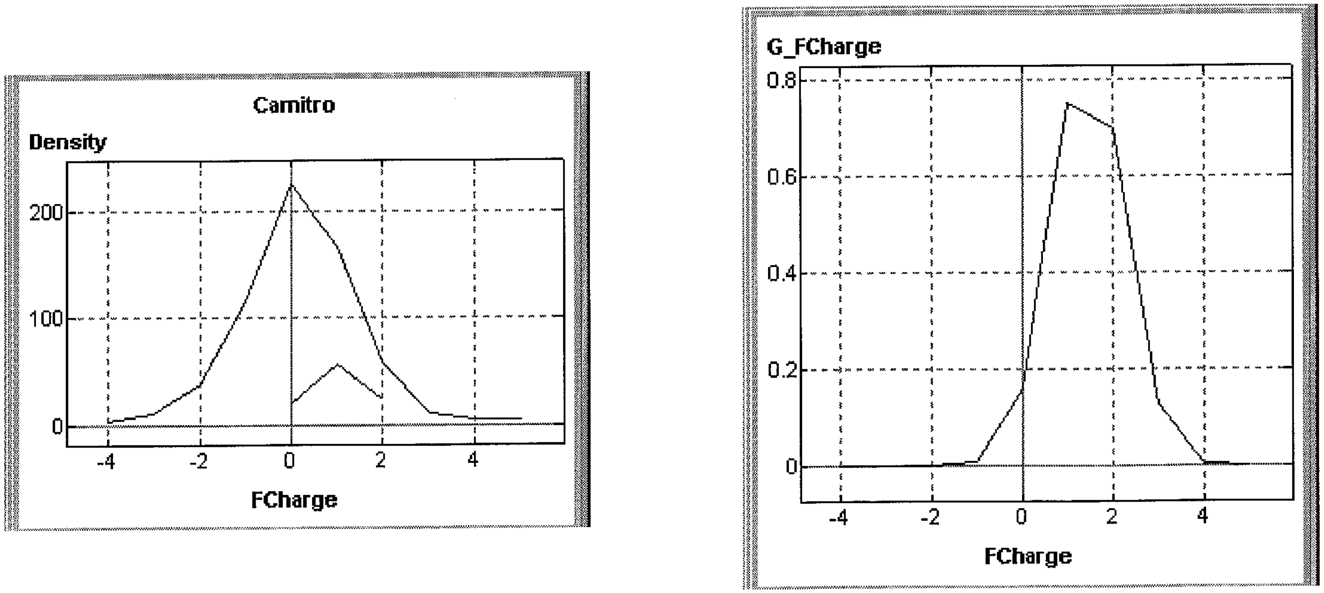
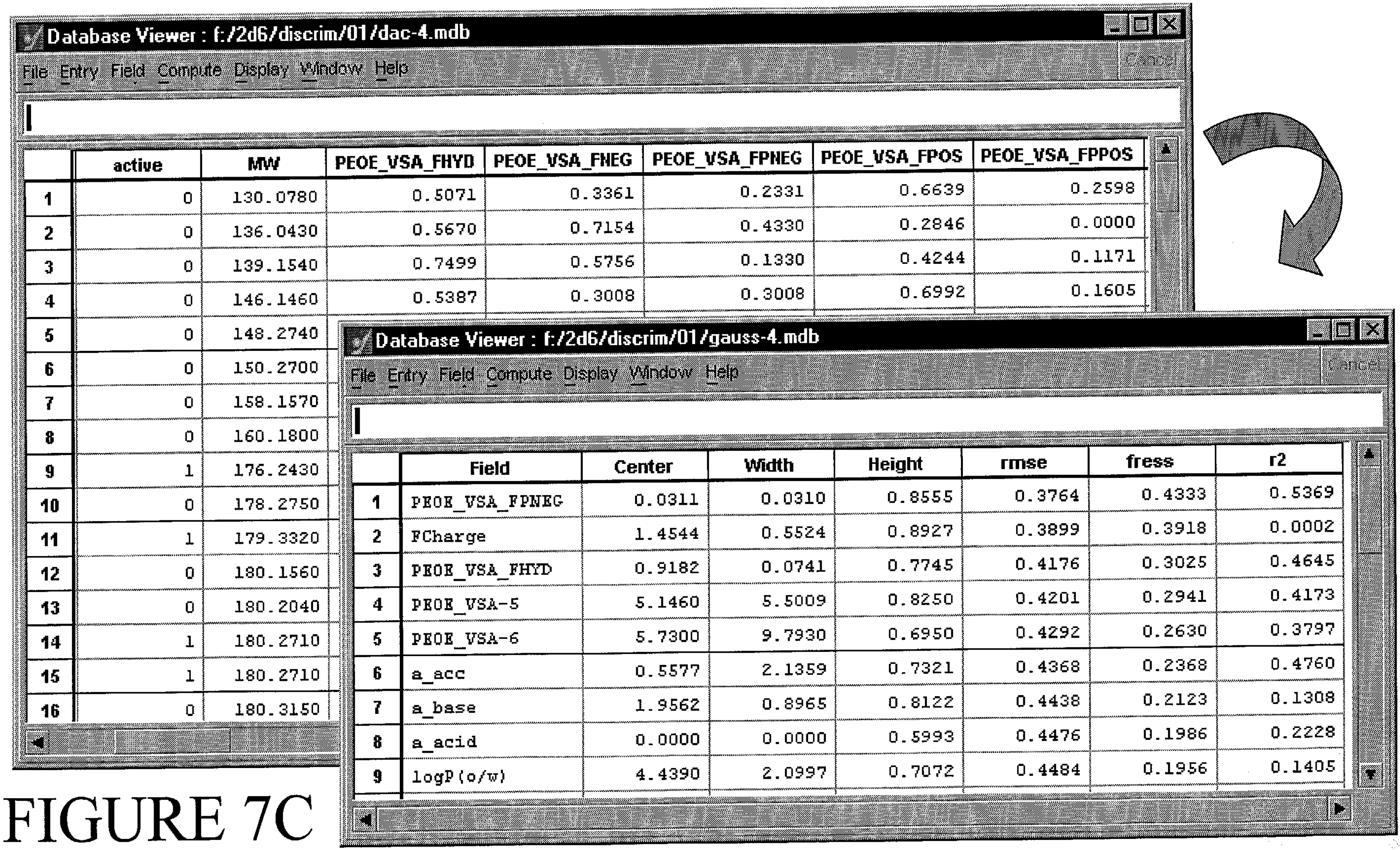


FIGURE 7B

Automated Gaussian Fit



	active	MW	PEOE_VSA_FHYD	PEOE_VSA_FNEG	PEOE_VSA_FPNEG	PEOE_VSA_FPOS	PEOE_VSA_FPPOS
1	0	130.0780	0.5071	0.3361	0.2331	0.6639	0.2598
2	0	136.0430	0.5670	0.7154	0.4330	0.2846	0.0000
3	0	139.1540	0.7499	0.5756	0.1330	0.4244	0.1171
4	0	146.1460	0.5387	0.3008	0.3008	0.6992	0.1605
5	0	148.2740					
6	0	150.2700					
7	0	158.1570					
8	0	160.1800					
9	1	176.2430					
10	0	178.2750					
11	1	179.3320					
12	0	180.1560					
13	0	180.2040					
14	1	180.2710					
15	1	180.2710					
16	0	180.3150					

	Field	Center	Width	Height	rmse	fress	r2
1	PEOE_VSA_FPNEG	0.0311	0.0310	0.8555	0.3764	0.4333	0.5369
2	FCharge	1.4544	0.5524	0.8927	0.3899	0.3918	0.0002
3	PEOE_VSA_FHYD	0.9182	0.0741	0.7745	0.4176	0.3025	0.4645
4	PEOE_VSA-5	5.1460	5.5009	0.8250	0.4201	0.2941	0.4173
5	PEOE_VSA-6	5.7300	9.7930	0.6950	0.4292	0.2630	0.3797
6	a_acc	0.5577	2.1359	0.7321	0.4368	0.2368	0.4760
7	a_base	1.9562	0.8965	0.8122	0.4438	0.2123	0.1308
8	a_acid	0.0000	0.0000	0.5993	0.4476	0.1986	0.2228
9	logP(o/w)	4.4390	2.0997	0.7072	0.4484	0.1956	0.1405

FIGURE 7C

2D6 K_i Model

Non-linear Size Relation

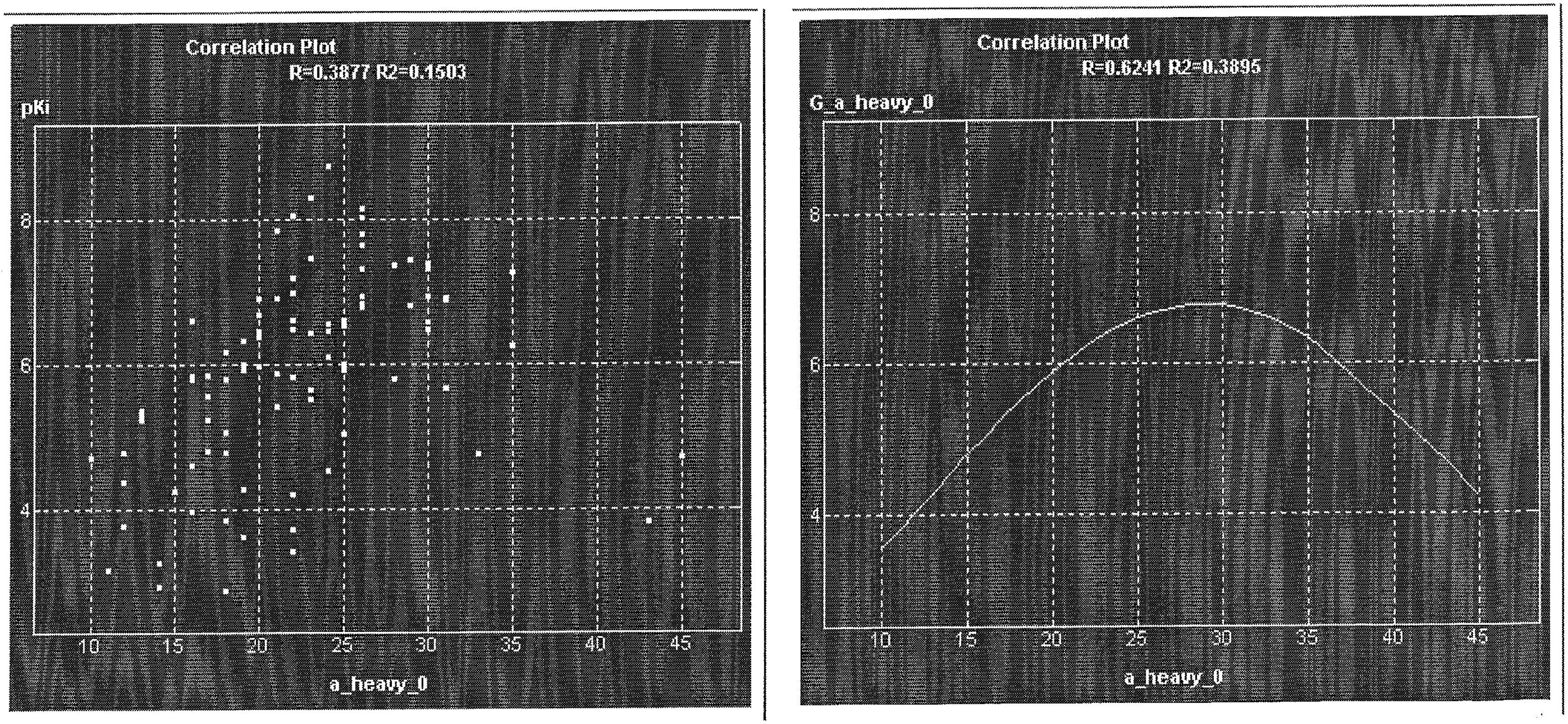


FIGURE 7D

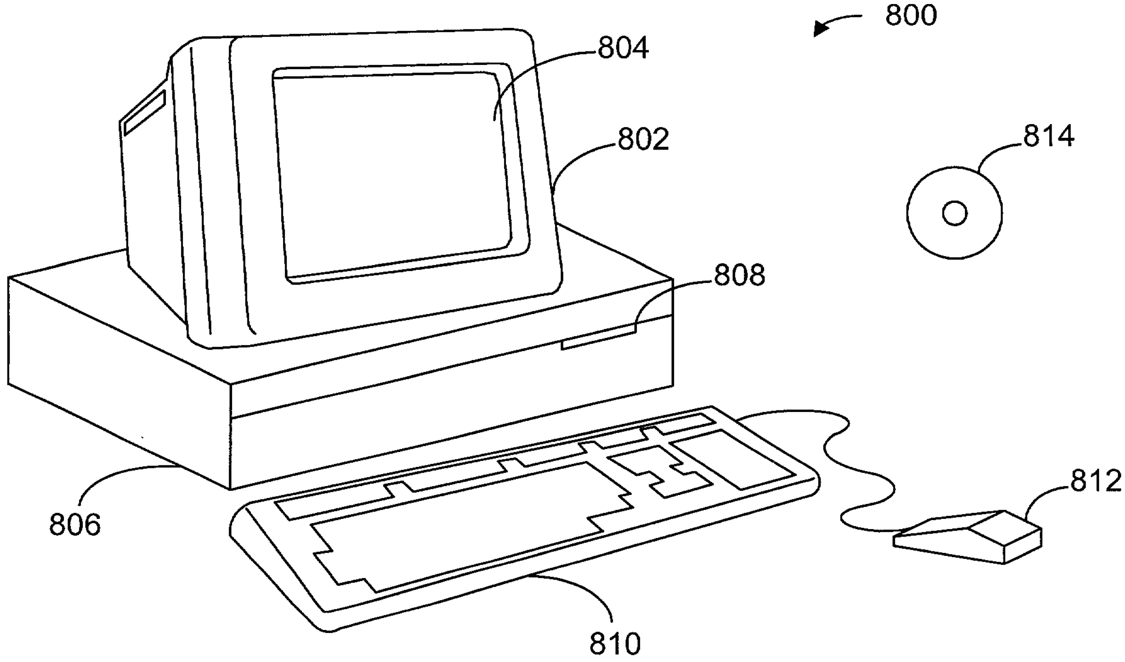


FIGURE 8A

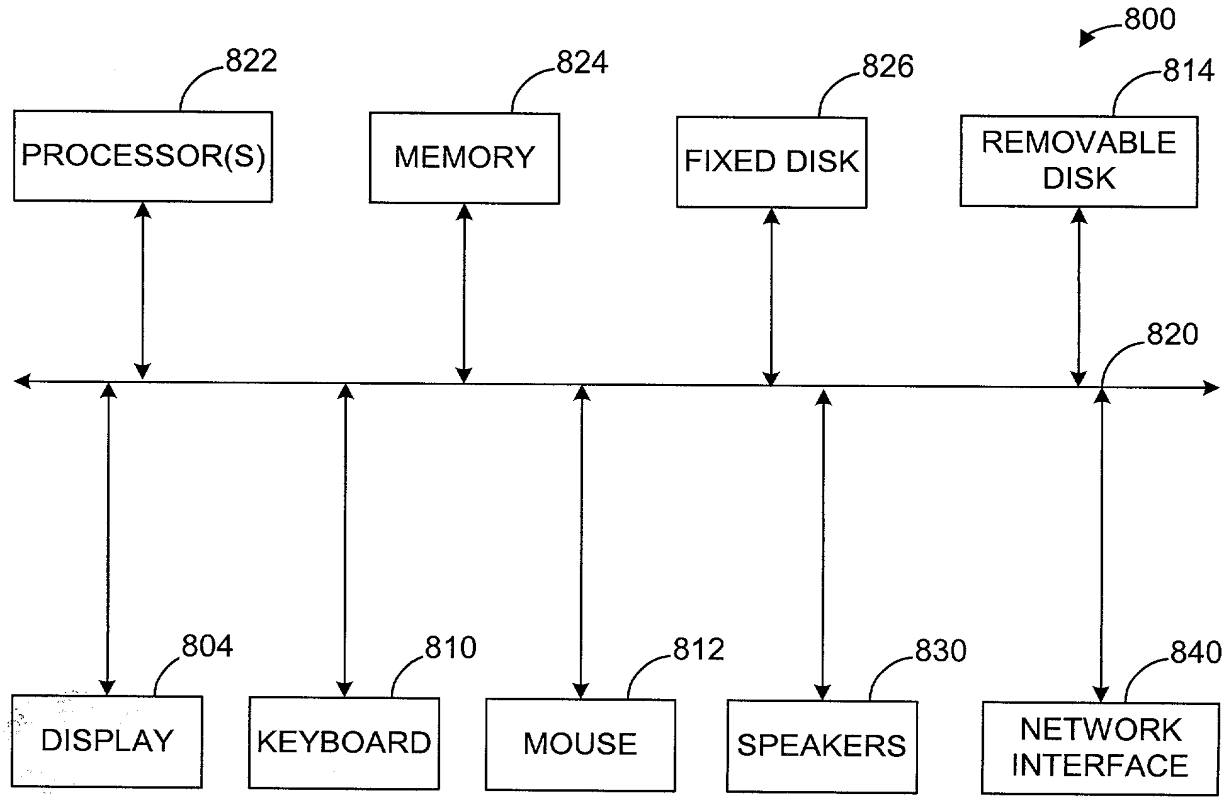


FIGURE 8B

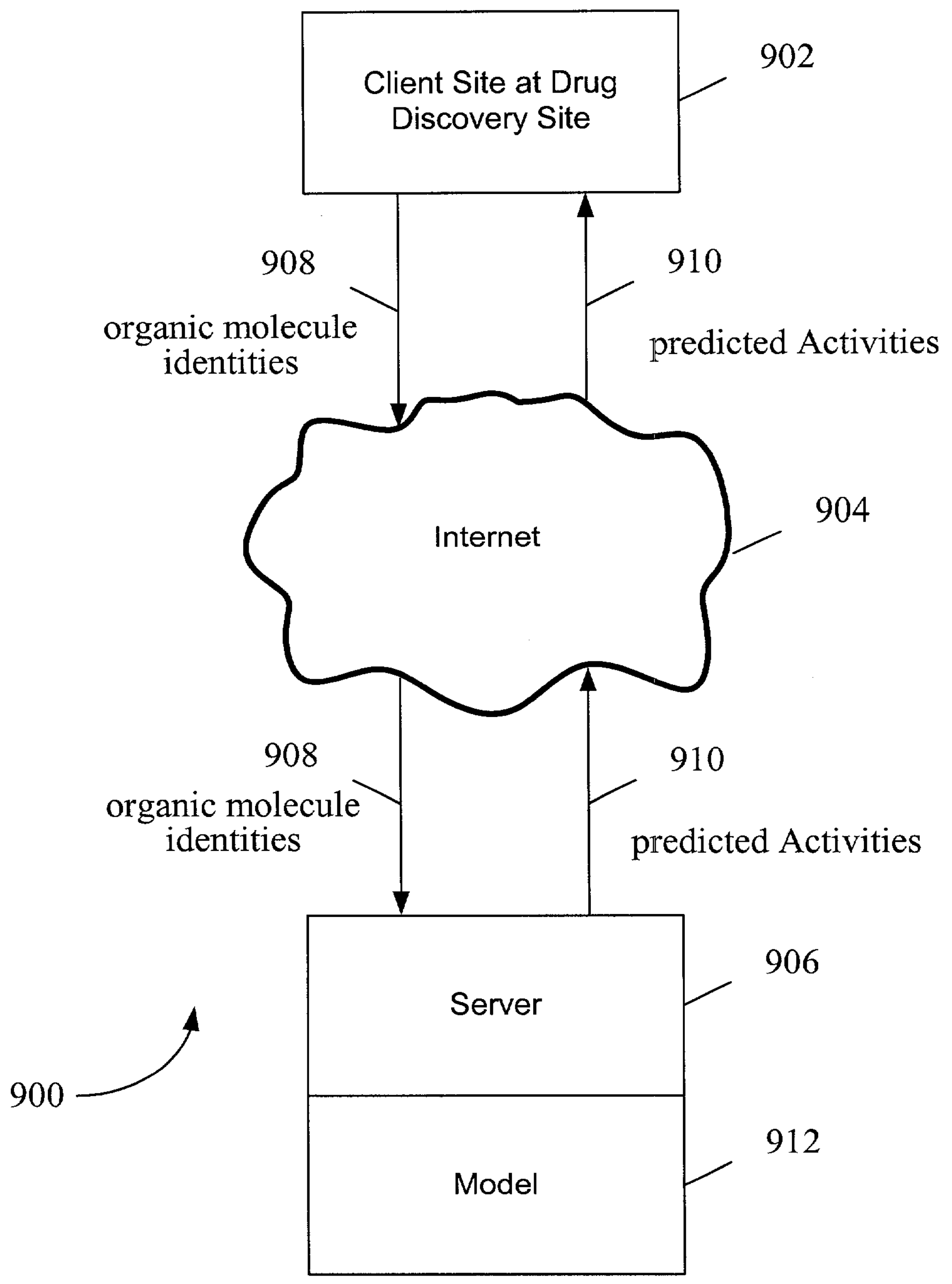

Client Site at Drug Discovery Site
902
908
organic molecule identities
910
predicted Activities
Internet
904
908
organic molecule identities
910
predicted Activities
Server
906
900
Model
912

FIGURE 9